DOCUMENT RESUME

ED 334 269 TM 016 913

AUTHOR Crews, William E., Jr.
TITLE Analysis of Interrater Reliability on the Evaluation of Answers to Open-Ended Questions.
SPONS AGENCY North Carolina State Univ., Raleigh. Center for Research in Mathematics and Science Education.; North Carolina State Univ., Raleigh. School of Education and Psychology.
PUB DATE Feb 91
NOTE 19p.; Paper presented at the Annual Meeting of the North Carolina Association for Research in Education (Chapel Hill, NC, February 28, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Elementary School Teachers; *Evaluation Methods; Evaluators; Grade 8; *Interrater Reliability; Junior High School Students; *Researchers; *Science Teachers; *Scoring; Testing Problems
IDENTIFIERS Halo Effect; *Open Ended Questions; Pearson Product Moment Correlation

ABSTRACT

As part of a study of teacher evaluation of student replies to open-ended questions, a second question--the best method of determining interrater reliability--was examined. The standard method, the Pearson Product-Moment correlation, overestimated the degree of match between researchers' and teachers' scoring of tests. The simpler percent agreement method tended to underestimate agreement. Scores were derived from two science teachers who were teaching a total of five eighth grade classes. Twenty students in each of the classes took the study tests at any one time. Two tests were used (the Earth and Moon Test and the Earth and Sun Test), which each contained 13 open-ended questions. The researcher and teachers evaluated the tests separately. Only the total number of correct answers for each student was used in the statistical analyses. Matched t-tests showed that most of the sets of scores could not be considered the same. A second sum of squares was derived fitted to the researcher's scores. A ratio of the two sums produced a fraction comparable to the correlation coefficient. Teachers usually evaluated answers adequately, but tended to count more answers correct than did the researcher for the higher ability classes. It is concluded that the halo effect was in operation; teachers should be warned to anticipate this type of bias. Four data tables are included. An appendix presents the instruments used. (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED334269

# Analysis of Interrater Reliability on the Evaluation of Answers to Open-Ended Questions

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

William E. Crews, Jr.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
William E. Crews, Jr.
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

paper presented at
the Annual Meeting of the
North Carolina Association for Research in Education

February 28, 1991

Chapel Hill, North Carolina

being part of a research project funded by
the Organized Research Fund
of the
School of Education and Psychology
and supported by the
Center for Research in Mathematics and Science Education
North Carolina State University
Raleigh, North Carolina 27695-7801

TM 016 913



BEST COPY AVAILABLE

Abstract

The question as to the feasibility of classroom teachers using open-ended questions to evaluate student conceptions was the initial problem. However, the standard method of calculating interrater reliaibility, Pearson's r, overestimated the degree of match between the researcher's and teachers' scoring of the tests because of systematic differences. Also, the comparison of percentage of scores that were the same seemed to underestimate the agreement where there were random differences between the scores. Therefore, a second objective became to find another method for expressing the degree of match.

Two teachers with five eighth-grade classes between them participated. There were two tests used with 13 open-ended questions each. The researcher and teachers evaluated the tests separately. Only the total number of correct answers for each student was used in the statistical analyses.

Matched t-tests showed that most of the sets of scores could not be considered the same. A second sum of squares was derived fitted to the researcher's scores. A ratio of the two sums produced a fraction comprable to the correlation coeficient.

In most cases the teachers adequately evaluated the answers. However, they tended to count more answers correct than the researcher did for the higher ability classes. Therefore, it would appear that the 'halo effect" was in operation and teachers should be warned to watch out for such bias in evaluating these types of tests.

Objective

This analysis was part of a pilot experiment. The question considered in this part of the experiment was, "Can classroom teachers reliably evaluate their students' replies to open-ended questions?" However, this article mainly discusses the answer to a question that occurred along the way, "What is the best method for determining interrater reliability?" This question became an issue when standard methods of analysis for interrater reliability and agreement produced discordant results with the differences in scores observed between the researcher and the classroom teachers. In order to answer the original research question, a method of representing interrater reliability that more accurately reflected the observed data needed to be determined.

Perspective

The standard method of measuring the match of two raters' scores is Pearson's Product-Moment correlation (r) (Gay, 1981). However, this method will produce a disproportionately high correlation between two sets of scores when there is a systematic error between the scorers. A simpler method of expressing interrater reliability is percent agreement (Gay, 1981). However, this method is too simple and would tend to underestimate the agreement between scorers in that it does not take into account the random variations in judgment between scorers.

The reason open-ended tests were being used and considered as an instrument for classroom teachers stems from the nature of the larger research project. The research being conducted dealt with identifying correct conceptions and misconceptions held by students concerning science. Open-ended questions are used in the development of the more recommended two-tiered, multiple-choice test for eliciting student

concepts (Treagust, 1988). Since much effort goes into the development and validation of such an instrument, the possibility of using the intermediate questions as an evaluation method by classroom teachers had been under investigation. Prior experiments had shown that the results obtained by open-ended questions closely matched the results obtained by interviews of students at the same age (Crews, 1990). Although the written evaluation tended to underestimate the correctness of the concepts held by students as compared to interviews, they were not significantly different (Crews, 1990). Interviews have generally been considered the most reliable means of eliciting student concepts (Treagust, 1988, and Bar, 1987). On the other hand, the two-tiered multiple-choice tests tend to produce very different results from interviews while written open- ended questions produce similar results to interviews (Bar, 1987).

Science teachers tend to give largely objective tests consisting of multiple-choice, matching and true-false questions. Some subjective, short-answer questions and problems may also be used. Therefore, a completely open-ended test differs from the normal test given and graded by a science teacher. On the other hand such open-ended tests tend to have fewer questions, so the time needed to evaluate a short answer by a classroom teacher could be offset by having fewer answers to check.

After consideration of these factors, the researcher hypothesized that an experienced and knowledgeable teacher should be able to adequately evaluate the correctness of student answers to such open-ended questions with a significant degree of accuracy. The degree of accuracy was defined as the researcher's evaluation of the same answers.

## Data Source

Two eighth-grade science teachers were involved in this evaluation. One of the teachers had four years of experience, but was teaching science for the first time. This teacher's subject-matter background was minimal. The other teacher was a first-year science teacher. However, this teacher's subject matter background was excellent.

The first teacher had three classes participating in the project. One class consisted of gifted and talented students from around the city who attended the magnet program offered at this middle school, although the science class was not specifically designated a GT class. The other two classes consisted of the remaining students who were mostly from ethnic, inner-city backgrounds.

The second teacher taught two science classes. These students were from rural and suburban neighborhoods outside the city. The teacher characterized one class as being smarter than the other. All classes had about 20 students (±3) take the tests at any one time.

## Methods

There were two tests consisting of thirteen open-ended questions each. The first test dealt with concepts concerning the Earth's relationship with the Sun. The second test contained questions about the Earth's relationship with its moon. The first test had been extensively field-tested and validated (Crews, 1990). The second test had been less extensively tested.

The teachers attended ten hours of instruction on the concepts covered in the units using the same methods which they were to use in their teaching of the two units. The teachers also took the tests for themselves and discussed the correctness of their answers.

The Earth and Sun test was administered as a pretest. For the gifted and talented class, the teacher needed to have an idea of the students' preconceptions, so was directed to grade their papers and record their evaluation on a grading sheet provided by the researcher. The researcher then evaluated all of the pretests from all five classes.

Between two and three weeks later, the teachers had finished instruction and were instructed to grade all tests in order to provide the students with grades while the researcher also evaluated them prior to their being given back to the students. The teachers were asked to record their evaluations on the question-by-question summary sheet provided (except for the GT class). However, one set of tests were returned without the teacher's grading being recorded.

At the same time the Earth and Sun posttest was given, the Earth and Moon pretest was completed by the students. Between two and three weeks later the Earth and Moon test was given again as the posttest. The teachers were asked to grade and record their evaluation of all of the answers on the posttest only. The researcher also rated the answers and recorded them on a grading sheet prior to the tests being returned to the students. At no time during the grading were the researcher and teachers aware of how each other had rated the answers.

Table 1. Mean Scores for Researcher's and Teachers' Evaluations

| | | Earth & Sun | | Earth & Moon | |
|---|---|---|---|---|---|
| | Class | Researcher | Teacher | Researcher | Teacher |
| Teacher 1 | 1st | 6.7 ± 2.8 | 5.9 ± 2.4 | 7.3 ± 2.4 | 9.6 ± 2.5 |
| | 2nd | 6.5 ± 1.6 | (N/A) | 4.5 ± 2.5 | 4.9 ± 2.3 |
| | 3rd | 6.8 ± 2.1 | 6.9 ± 2.1 | 4.6 ± 2.9 | 5.4 ± 3.2 |
| Teacher 2 | 1st | 9.9 ± 2.3 | 10.0 ± 2.1 | 6.4 ± 2.2 | 6.9 ± 2.3 |
| | 2nd | 8.0 ± 3.3 | 8.1 ± 3.2 | 4.5 ± 2.8 | 4.9 ± 2.6 |

## Results

The teachers' scores were generally close to those of the researcher (see Table 1). There was a difference of -.8 ±1.0 between the teacher's scoring on the Earth and Sun pretest and the researcher's. There was much less of a difference on the posttest (+.2 ±1.3) with the teacher commenting that it was much easier to evaluate the answers after having taught the unit. The other teacher with the two classes had slightly higher average differences than the researcher (+.4 ± 0.9 and +.3 ± 1.2).

There were more widely varying differences between the two sets of scores on the Earth and Moon posttest. On the tests for the gifted and talented class, the first teacher produced scores an average of 2.3 ± 1.1 points higher. For the less capable classes, the differences were a little less, but more variable at +.4 ± 1.5 for the second class and +.8 ± 1.2 for the third class. The second teacher had average differences of +.5 ± 1.3 and +.4 ± 1.3 on the two classes participating in the project.

Interrater Reliability Calculations

Conducting regression analyses and computing Pearson's r for each set of scores produced very high correlations (see Table 2). According to these results, the sets of scores were very similar, much more than could ever be expected to have occurred by chance. However, examination of the two sets of scores on a pupil-by-pupil basis showed some systematic differences in scoring on the Earth and Moon posttest where the teacher had consistently counted certain answers correct that the researcher counted wrong. In order to see if the two sets of scores could be considered the same, matched t-tests were calculated for each ).

Table 2. Pearson's Product-Moment Correlations for the Evaluations

| | | Earth & Sun Tests | | Earth & Moon Tests | |
|---|---|---|---|---|---|
| | Class | r | probability | r | probability |
| | 1st | 0.93 | <<0.001 | 0.89 | <0.001 |
| Teacher 1 | 2nd | N/A | 0.82 | <0.001 | |
| | 3rd | 0.67 | <0.001 | 0.93 | <<0.001 |
| | 1st | 0.94 | <<0.001 | 0.84 | <0.001 |
| Teacher 2 | 2nd | 0.93 | <<0.001 | 0.89 | <0.001 |

The matched t-tests produced very different results from the correlations. In the instances where the average differences were higher, the probability of the two sets of scores being the same was equal to or lower than the standard significance level of 0.05. In the instances where the average difference between the researcher's and teachers' scores were lower, the matched t-scores suggest that the two sets of scores were, indeed, equivalent.

Table 3. Matched t-scores for the Researcher's and Teachers' Evaluations

| | | Earth & Sun Tests | | Earth & Moon Tests | |
|---|---|---|---|---|---|
| | Class | t | probability | t | probability |
| Teacher 1 | 1st | 3.3 | 0.004 | 8.7 | <0.001 |
| | 2nd | N/A | | 1.1 | 0.30 |
| | 3rd | 0.66 | 0.50 | 3.2 | 0.006 |
| Teacher 2 | 1st | 2.3 | 0.03 | 2.1 | 0 05 |
| | 2nd | 1.3 | 0.20 | 1.7 | 0.10 |

Because of the disparity between the two analyses, another method of comparison was sought. In consultation with a statistician, it was suggested that the least sums of squares from the standard regression analysis used to calculate Pearson's r be compared to a larger sums of squares produced from a regression line where the researcher's scores were considered to be the correct scores (X's). This ratio produces a number

comprable to Pearson's r. In addition an educational researcher who was consulted for help on determining a better method of measuring interrater reliability suggested that the r[illegible] of the two sums of squares could be reversed and would prod[illegible] statistic of n,n degrees of freedom from which a probability level cou[illegible] ooked up on a standard F table. The results of this analysis can be [illegible] Table 4.

Table 4. Correlations and P[illegible]es from Comparison of Sums of Squares

| | | Earth & Sun Tests | | Earth & Moon Tests | |
|---|---|---|---|---|---|
| | Class | correlation | probability | correlation | probability |
| Teacher 1 | 1st | 0.62 | ≈0.15 | 0.18 | <0.005 |
| | 2nd | N/A | | 0.74 | ≈0.15 |
| | 3rd | 0.93 | ≈0.15 | 0.67 | ≈0.15 |
| Teacher 2 | 1st | 0.61 | ≈0.11 | 0.81 | ≈0.18 |
| | 2nd | 0.97 | ≈0.20 | 0.80 | ≈0.18 |

These correlations and matching probabilities more closely align with the results of the matched t-tests. Only one set of scores produced an insignificant probability, the Earth and Moon test for the gifted and talented class which also had the highest mean difference between teacher and researcher scoring.

## Conclusions

There was an initial question that this data was collected to answer, "Could teachers reliably score these open-ended tests?" The second question, though, had to be answered before the first one could be and that was, "What is the best way to statistically measure interrater reliability?" This question is answered in the final section.

In general it was concluded that the teachers could do a satisfactory job of correctly evaluating the answers. However, the one case of a significant difference that occurred in the Earth and Moon posttest for the GT class pointed out a problem in the teacher evaluation that could also be seen, to a lesser extent, in the grading of the other tests and classes. Looking at the individual scores for the GT students, it could be seen that the teacher tended to consistently check many answers as completely correct when the researcher did not consider them so. Apparently the teacher was more willing to give the smarter students the benefit of the doubt when evaluating their answers. A similar, but not as significant, trend could be seen in the teacher with the two classes. The class previously identified as containing smarter students had slightly higher average difference scores between the researcher and teacher.

## Educational and Research Implications

The "halo effect" is well documented and was seemingly at work in this experiment. Therefore, when using open-ended questions as a classroom evaluation tool, teachers should be warned about subconscious biases and be sure they are evaluating answers based solely on what is written down rather than include their impressions of how much they think the student knows because these teacher impressions are not always reliable.

A comparison of the three methods of analyses along with the general conclusion gained by the researcher from comparing the individual scores found the most informative method of expressing interrater agreement to be the sums of squares correlation. The probability of match between the scores seemed too high and susceptible to systematic errors when using Pearson's *r* for analysis. The matched t-tests were useful in determining that the scores could or could not be considered the same, but did not give an idea as to the degree of match of the scores. Considering the *t*-scores as fairly rigorous, it can be seen that the correlation coefficients and associated probabilities derived from the sums of squares comparisons closely follow the *t*-score probabilities with minor differences attributable to differences in degrees of freedom.

A disadvantage of this method is the necessity for an expert's ranking. The researcher's ranking was used as the basis for comparison during the normal regression analysis with another rater's scores to produce the least squares sum. This type of comparison does not recognize the researcher's rating as necessarily being the most correct, but calculates an average X based on all of the information in the regression. However, the researcher's rating was considered as correct and fixed indpendent variable measurements in the fitted regression to produce the slightly larger second sum of squares from which the ratio and *F* statistic is calculated. If there is no set of standard scores which can be used for this expert comparison, the sums of squares method for calculating a correlation coefficient and matching probability will not work.

References

Bar, V. (1987). The effect of the testing format on the distribution of results. In Novak, J. D. (Ed.), Proceedings of the Second International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Volume I (pp. 26-31). Cornell University, New York.

Crews, W. E. (1990). Development of a paper-and-pencil instrument for assessing student conceptions concerning the Earth as a planet. Paper presented at the National Association of Researchers in Science Teaching annual meting, Atlanta, Georgia, April.

Gay, L. R. (1981). Educational Research: Competencies for Analysis & Application (2nd ed.). Columbus, OH: Charles E. Merrill.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. International Journal of Science Education, 10(2), 159-169

# APPENDIX

## The Instruments

Name: _____ Class Period: _____ Date: _____
Teacher: _____ School: _____

## EARTH AND MOON TEST

Answer the following questions as best you can. Answer each one with a full sentence. **For example:** Where is the Moon in relation to the Earth?
*The Moon is far away from the Earth in outer space, but not as far as the Sun.*

1. Name and/or draw the four main phases of the Moon.

_____ _____

_____ _____

2. In what order do these four phases occur?

_____

_____

_____

_____

3. How long does it take for the Moon to go though all of its phases once?

_____

4. What causes the Moon to go through these phases that we observe?

_____
_____
_____

5. Is the Moon seen in the sky every night (when it is not cloudy)? _____

b. Why or why not? _____
_____

6. Is the Moon seen all night long in the sky (when it can be seen)? _____

b. Why or why not? _____
_____

7. Is the Moon ever seen in the sky during broad daylight? _____

   b. Why or why not? _____

_____

8. Why do we see the same side (or a part of that side) of the Moon all the time when we see the Moon?

_____

_____

_____

9. How many times does the Moon go through all of its phases in a year?

_____

10. What is a Solar Eclipse? _____

_____

11. Why are there only one or two Solar Eclipses per year around the world?

_____

_____

_____

12. Do Solar Eclipses occur during any particular phase of the Moon? _____

   b. If so, which one? _____

13. What is a Lunar Eclipse? _____

_____

14. Why are there only one or two Lunar Eclipses per year around the world?

_____

_____

15. Do Lunar Eclipses occur during any particular phase of the Moon? _____

   b. If so, which one? _____

Name: _____ Class Period: _____ Date: _____
Teacher: _____ School: _____

## EARTH AND SUN TEST

Answer the following questions as best you can. Answer each one with a full sentence. **For example:** Where is the Sun in relation to the Earth?
*The Sun is far away from the Earth in outer space.*

1. Describe the shape of the Earth. Use an example to make your answer clear.

_____
_____

2. What is the shape of the Sun? Also use an example of something else that is the same shape to make your answer clear.

_____

3. During the day, the Sun can be seen to rise, move across the sky, and set. What is causing this movement that we see?

_____
_____
_____

4. What happens so that we are not able to see the Sun at night?

_____
_____

5. Most people who live on the East Coast have watched football or baseball games broadcast from California. If you have done so, then you may have noticed that there is still daylight on the West Coast for about three hours after the Sun has already set here in North Carolina. Why is that?

_____
_____

6. You have been answering questions about the way the Earth and Sun move that causes day and night. How do the Earth and Sun move in relation to each other that causes the changing cycle of the year?

_____
_____
_____

NAME: _____

Most people have noticed that there are some times during the year when we have more hours of daylight and other times when we have fewer hours of daylight than darkness during a day.

7. During what season do we have the most hours of daylight? _____

   b. Why is there more daylight during this part of the year?

_____
_____
_____

8. During what season do we have the least hours of daylight? _____

   d. Why is there less daylight during this part of the year?

_____
_____
_____

9. There are also two seasons when there is almost exactly the same number of hours of daylight and darkness in the day. When are these?

   a. _____ b. _____

10. What causes the hours of daylight and darkness to be about even during these two seasons?

_____
_____
_____

11. Of the Earth, Sun and Moon, which is smallest? _____

    b. Which is largest? _____

12. Which is closer to us (the Earth), the Sun or Moon? _____